Visual Prosthesis: Enhancing Daily Experiences for People with Visual Impairments

Yumeng Ma
Brown University
Providence, USA
yumeng_ma1@brown.edu

ABSTRACT

We introduce a visual prosthesis system for enhancing the mobility and independence of individuals with visual impairments. By combining object localization and optical character recognition within a wearable framework, we offer a solution that prioritizes intuitive use and social discretion. The system consists of glasses, equipped with a camera and auditory feedback mechanisms, to facilitate real-time environmental awareness and text interaction. Users control the device through voice commands to identify objects and read text in their surroundings without the need for physical interaction with the device. This approach reduces the social stigma associated with assistive devices and improves the quality of interaction and exploration with their environment. Our development process emphasizes user feedback and practical trials to refine functionality and ergonomics. Future enhancements will focus on expanding the system's capabilities to include direct object manipulation and navigational enhancements to accommodate a wider range of daily activities and settings.

Author Keywords

People with visual impairments, Object detection, Object localization, Optical character recognition

INTRODUCTION

Navigating the world presents challenges for people with visual impairments (PVI). Although current assistive technologies offer some aid, they often fall short of addressing the needs of PVI. Common issues with these devices include their intrusive and conspicuous nature, high costs, and unreliable performance, which can intensify feelings of dependency and self-consciousness among users [5].

While efforts have been made in computer vision, such as real-time object detection systems like YOLOv5 [25] and optical character recognition (OCR) tools including EasyOCR [2] and Tesseract [22], their application in assistive devices for PVI has not been investigated enough. Moreover, they are often not designed with the user experience of visually impaired individuals in mind.

Recognizing these gaps, we developed a visual prosthesis system that refines these tools to specifically address the needs of blind and low-vision users. Our system consists of a pair of glasses equipped with a camera and headphones featuring a microphone. We enhance autonomy and interaction by combining object localization and OCR into a cohesive unit catered for accessibility. In our version:

- ***Object localization*** leverages current object detection and allows users to verbally inquire about their surroundings. The system responds through audio to inform the user of the types and locations of objects within their environment. This information consists of both the direction (using a clock-face reference) and the distance to the objects for spatial orientation. For instance, a user might say, *"find"* and receive a response like, *"Chair, at three o'clock, at 1.5 feet."*
- ***OCR*** utilizes existing OCR capabilities that can be activated by issuing a voice command like *"t"* to capture and process text from the environment. The recognized text is then read aloud using text-to-speech. Users can navigate through the text with commands like *"n"* to read the next line or finish saying the sentence *"s"*. We also make adjustments to the text grouping and reading order and implement horizontal text correction for accommodating skewed textual formats encountered in real-world settings.

We challenge existing norms and introduce a new standard for what assistive technologies can accomplish. By prioritizing user empowerment over assistance, we see how such a system can serve as an extension of personal capability and enable PVI to navigate the world with more confidence.

RELATED WORKS

Existing Technologies for PVI

Smart glasses and wearables have made iterative advancements geared towards helping those who are blind and low vision. Devices like OrCam MyEye [24], which clips onto glasses to read text aloud and recognize faces, and eSight [9], which uses high-definition cameras to enhance visual information for PVI, represent growth in wearable technology. However, these solutions are often criticized for their high costs, low battery life, and steep learning curve. Research has also shown that the stigma associated with their conspicuous designs can alienate users from their social environments [19].

Additionally, products like Ray-Ban's smart glasses [17] often struggle with camera field of view limitations; their cameras typically do not provide a sufficiently wide or appropriately oriented field of view, making them less effective for spatial awareness. Many augmented reality (AR) glasses face similar challenges, with fields of view that are either too narrow or poorly aligned (too vertical or too horizontal) [23]. This limitation affects their utility in 'out-in-the wild' environments where dynamic interaction is necessary such as navigating through crowded spaces or adapting to varied outdoor landscapes. Such scenarios demand a panoramic view that most current AR glasses fail to deliver.

Mobile applications have also developed to support PVI in daily tasks. Microsoft's Seeing AI app [11] provides spoken feedback to describe visual elements such as text, faces, and scenes by leveraging the smartphone's camera. Similarly, Be My Eyes [10] connects visually impaired users with sighted volunteers who assist them through live video calls. Despite their utility, these apps often rely on crowdsourcing and continuous internet connectivity, which can limit their effectiveness in more rural or remote areas, where internet service is inconsistent or unavailable. Further, individuals from socioeconomically disadvantaged backgrounds might not have regular access to reliable internet services, making these tools less accessible. This accessibility gap can disproportionately affect underrepresented groups who might already face additional barriers to technology usage.

Object Detection and OCR for Accessibility

Many object detection and OCR applications are predominantly designed for use on smartphones or tablets. This means that users manually capture the environment around them. Popular examples include object detection apps like Google Lens [13] and CanFind [15] and text scanners with OCR capabilities like Adobe Scan [1] and Google Keep [12]. The use of these apps often requires active user engagement, making the use of assistive devices intentional and deliberate. For instance, using a smartphone to read signage or identify objects on a busy street typically involves pulling out the device, holding it up, and pointing it in a targeted direction. This can draw attention and disrupt daily routine. These applications often operate on a request-response model, where the user must physically prompt an action to receive information. The lack of spontaneity can hinder the fluidity with which PVI can navigate or interact with their environment.

Studies from the VizWiz project [4] have delved deeper into refining object detection and OCR catered for PVI. While their dataset address challenges, such as handling images with large, simple-boundary salient objects or those lacking textual content [17], which are common in the real-world scenarios faced by PVI, their system still relies on the user initiating actions like taking pictures with a device and does not encourage passive exploration of the environment. These insights underscore the need for more adaptive solutions.

Our proposed system seeks to mitigate these issues by embedding object detection and OCR capabilities within a wearable format that operates as summoned. This approach ensures that the technology serves as an unobtrusive extension of the user's senses, providing real-time, contextual information through audio feedback without the need for conspicuous manual scanning.

SYSTEM DESIGN

The visual prosthesis system is designed to enable blind and low vision individuals to explore their surroundings and improve their contextual awareness.

Hardware

The system integrates into the user's environment through a wearable device, consisting of:

- *Glasses:* 3D printed frames equipped with a Luxonis [16] camera that captures the visual field in front of the user shown in Figure 1. The setup also includes an LCD overlay with adjustable opacity for testing purposes.
- *Headphones with Microphone:* Headset shown in Figure 2 for receiving voice commands from the user and delivering auditory feedback via text-to-speech synthesis.

Software

The interaction with the system is voice-based, using text-to-speech OpenAI's Whisper speech-to-text [20] for command recognition. Users can issue commands to the system to receive different types of feedback (note that it also supports keyboard shortcut inputs). The system's architecture is split into two modules:

Object localization uses YOLOv5 to detect a specific set of everyday use objects. Currently, it detects the classes from COCO dataset [6]. This mode recognizes objects within the user's vicinity and provides detailed spatial information: the object's clock position and its distance (in meters or feet) relative to the user depicted in Figure 3.

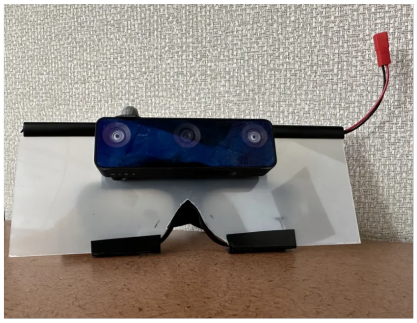
There are three commands that are supported by this module.

- *"Find" Command*: Enumerates objects detected in the scene from left to right, providing depth information for each identified object.
- *"List" Command*: Offers a quick summary of detected objects without depth information, useful for understanding the general composition of the scene.
- *Object-Specific Query*: Allows users to inquire about specific objects (e.g., *"orange"*). The system responds with the location and distance of the object if available, or a notification of absence, stating *"sorry, I can't locate that"*, if the object is not detected or recognized.

Figure 4 demonstrates examples of how each of these commands operate within the system.

OCR is operated through voice commands. We have improved text grouping and reading order to organize related text elements so that users can understand advanced textual layout like menus, signs, or documents with more ease. We have incorporated text correction as seen in Figure 5 which first converts text images to grayscale to simplify the data. It then applies Gaussian blur to smooth out the image and reduce

(a). 3D printed glasses frame



(b). Front view of glasses

(c). Back view of glasses

Figure 1. Setup of the visual prosthesis system's glasses. The camera is mounted on the top of the glasses frame, and the translucent LCD overlay is placed in front of the glasses.

Figure 2. Set up of the headset with an integrated microphone

noise and detail. Finally, it uses edge detection followed by Hough lines transformation to correct texts that are skewed or misaligned. The architecture of this module is depicted in Figure 6.

- *"Take Photo (T)"*: Saying "T" activates the camera to capture the text area in front of the user.
- *"Next Sentence (N)"*: Saying "N" moves to the next sentence in the OCR result.
- *"Finish Reading (S)"*: Saying "S" stops the reading process after it finishes the sentnece.

Figure 7 illustrates an example scenario where these commands are consecutively used.

DISCUSSION

Our system differs from traditional assistive devices by focusing on integration into the everyday lives of PVI. We make two fundamental contributions:

1. *Redefining Accessibility.* We propose a shift from assistive to accessible to show how technology can support and enhance natural abilities rather than compensating for disability. This shift is reflected in our design, which ensures the system is non-intrusive and integrates smoothly into the personal space of the user to foster a sense of normalcy and independence.
2. *Holistic Integration.* By synthesizing object detection and OCR capabilities into a cohesive framework, our system addresses a wide array of visual tasks with a single, unified device. This approach simplifies the user experience and widens the functional capacity of the system so that PVI can interact naturally with their environment.

The principle of ability-based design advocates for shifting the focus from requiring users to adapt to rigid computer systems to instead modifying these systems to accommodate the abilities of the users [26]. This approach underpins our objective of enhancing autonomy and accessibility for PVI. By emphasizing their capabilities rather than their limitations, we can foster greater independence in their daily activities through a system that is supportive yet unobtrusive and free from stigma.

During the design process, we encountered several challenges that tested our approaches and illuminated potential areas for improvement.

Limitations

Hardware Design. We recognize the need to fundamentally rethink our wearable setup. One of the primary concerns with the current design of our glasses is the slight bulkiness due to the Luxonis camera. This affects both the aesthetic appeal and the comfort of the device, which remains somewhat conspicuous and potentially uncomfortable during prolonged use. Our goal is to refine or find a design that resembles the sleek form factor of Ray-Ban glasses but incorporates a high-quality, more wide-angled camera that surpasses the capabilities of similar consumer products. Achieving this would better satisfy the discretion of the device. Inspired by recent advancements in AR glasses, we aim to embed speakers directly to the glasses (the ones that are capable of transmitting audio to the user without being overheard by bystanders). Additionally, we are working to eliminate the wires in our current setup to further enhance the usability and comfort of our system.

Object Detection. Object Detection: To enhance the reliability of our object detection, we are seeking to improve or find a superior dataset. Occasionally, the current system misidentifies objects, prompting us to search for a more accurate and de-

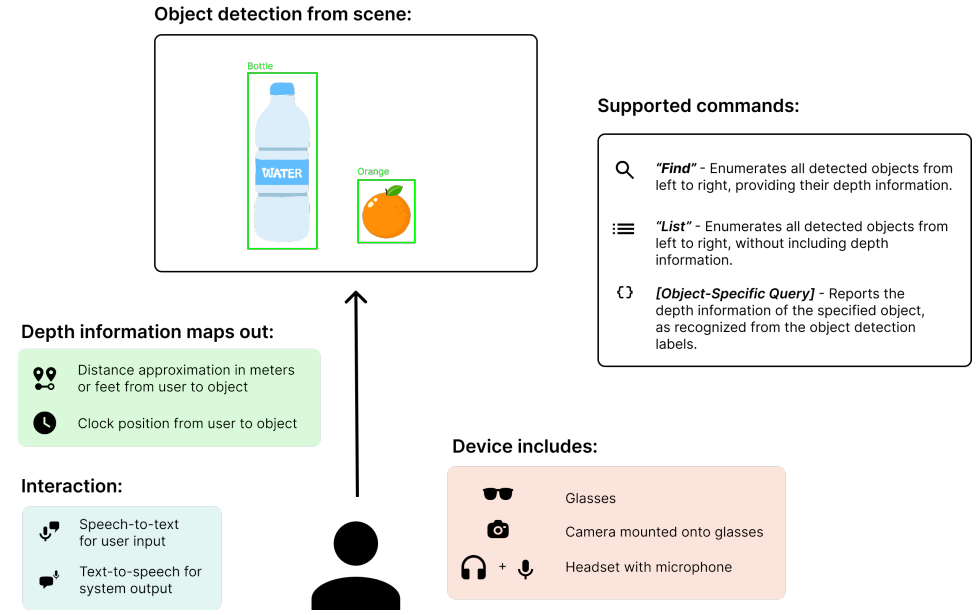
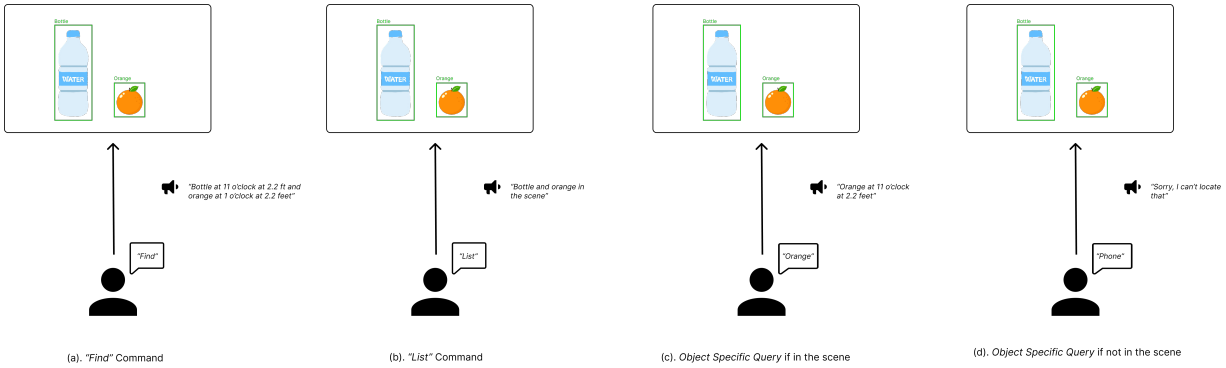


Figure 3. Architecture of the object localization module using YOLOv5 for real-time object detection. The system receives input from the scene and performs object recognition to identify objects within the user's vicinity. The user interacts with this module through voice commands to inquire about objects which triggers text-to-speech responses to provide depth information about the object.



(a). *"Find"* Command (b). *"List"* Command (c). *Object Specific Query* if in the scene (d). *Object Specific Query* if not in the scene

Figure 4. Illustration of an example responses of the system to the *"Find"*, *"List"*, and *"Object-Specific Query"* commands when an orange and a water bottle is in front of the user.

(a). Default image captured (b). Postprocess result

Figure 5. Comparison of text correction results before and after processing using an image captured by the Luxonis camera.

pendable solution, such as the VizWiz-SalientObject Dataset [21].

Speech Recognition. We experimented with a hot word detection that uses the trigger word *"computer"* to activate the object localization commands. However, the necessity for users to repeat this trigger word each time they wish to issue a voice command makes the system less natural. The speech-to-text functionality powered by OpenAI's Whisper also requires further tuning, especially in noisy settings. The variability in performance can lead to user frustration when voice commands do not result in accurate responses. To address this, we plan to delve deeper into the configurable parameters of Whisper's model in compatibility with our current headset and possibly experimenting with additional preprocessing techniques to enhance audio clarity.

Text Correction in OCR. Our current OCR setup involves a sequence of preprocessing steps to correct misaligned text. Initially, the text image is converted to grayscale, followed

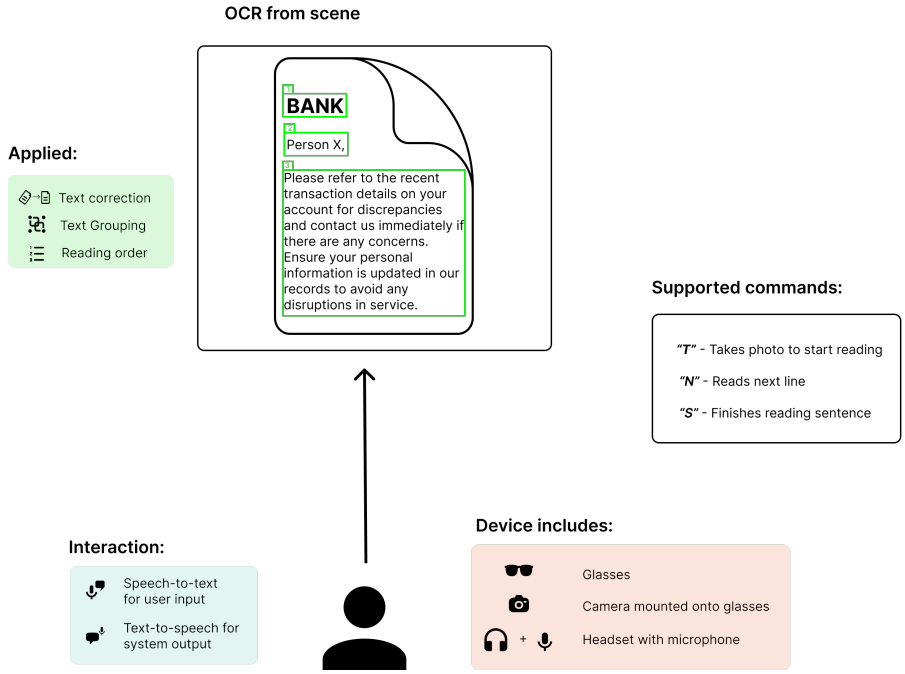


Figure 6. Architecture of the OCR module using EasyOCR. The system processes text input from the environment by applying text correction, text grouping, and reading order. The user interacts with the module through voice commands that control how and when the text is read to them.

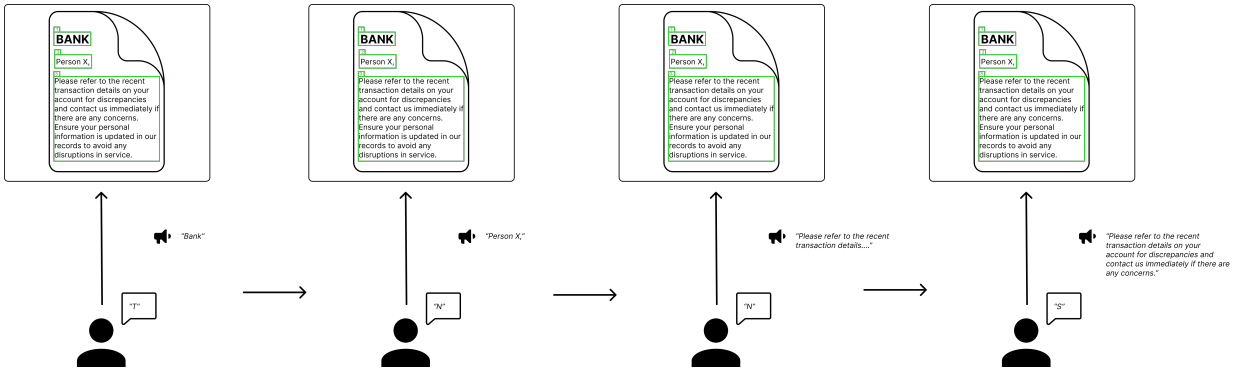


Figure 7. Illustration of a sequential use of the OCR module commands in a user scenario. The user initiates the process by saying *"T"* to capture and start reading the detected text. Subsequently, *"N"* is used to advance through the text, reading one line at a time. Finally, *"S"* is commanded to stop the reading process after completing the first sentence.

by the application of a Gaussian blur to reduce noise. We then perform edge detection to highlight the structure of the text. After these steps, we use Hough line transformation to identify and align the text lines properly. We noticed that this method still result in errors with more extreme cases when text is skewed. To enhance the accuracy of our text correction, we are interested in exploring an iterative correction process. This proposed method would involve repeatedly applying OCR to progressively tune the alignment and readability of text.

Future Works

Object Grasping Module. We are developing a third module focused on object grasping so that users can physically engage with their environment. The module utilizes the YOLOv5 model like the object localization to accurately detect and localize objects along with MediaPipe [14] to detect and track the position and movement of the user's hand in real time. We then want to use text-to-speech to guide the hand of the patient to touch and grasp the desired object of the specific set of everyday use objects. This module would guide the user's hand towards the object using directional sound cues such as *"up"*, *"down"*, *"left"*, and *"right"* to lead the user's hand to a designated target.

Connectivity. Our system's design intentionally avoids reliance on internet connectivity to ensure it can be used in any environment. This decision limits the system's ability to utilize other powerful cloud-based services such as Google's text-to-speech or speech-to-text. We plan to explore the potential of embedding more sophisticated processing approaches into the device, but this might include localized versions of services that typically require online connectivity. Our code currently runs on the NVIDIA Jetson [7]. Moving forward, we plan to develop a mobile app that pairs with our wearable device and

transition our system into an off-the-shelf product suitable for everyday use.

Adding Sensory. GPS and depth sensors in future models could be another avenue. These additions will provide better visual salience to detect obstacles such as stairs and provide more detailed environmental interactions.

Use of Large Language Models. Advancements in large language models (LLMs) and image captioning have proven effective in improving contextual interactions between computers and human language [27, 3, 18, 8]. Integrating these components into our system will give us the opportunity for users to make dynamic conversational inquiries about their surroundings that can be answered with the targeted information extracted. A user could direct the OCR to focus on particular types of information within their visual field. For instance, they might ask the system to identify and read only restaurant names or to provide detailed information about the contents of an aisle in a grocery store. The user could even request summarized descriptions of their surroundings.

As we advance, our commitment deepens to crafting a system that becomes an empowerment for PVI. We want to facilitate PVI's engagement with the world in ways that encourage exploration and independence. Heightening the intersection of human capability and technological advancements propels our research development efforts toward designing for its utility and reinforces our core belief that technology should conform to human needs and be inclusive by design.

ACKNOWLEDGMENTS

I thank my collaborators: Jorge Chang, for the substantial contributions to the system's codebase and consistent mentorship throughout the development process; and Michael Paradiso, for introducing me to this project and providing ongoing guidance and support.

REFERENCES

[1] Adobe. 2023. Adobe Scan. https://www.adobe.com/acrobat/mobile/scanner-app.html. (2023).

[2] Jaided AI. 2023. EasyOCR. https://github.com/JaidedAI/EasyOCR. (2023).

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[4] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and others. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 333–342.

[5] Shiwei Chen, Dayue Yao, Huiliang Cao, and Chong Shen. 2019. A novel approach to wearable image recognition systems to aid visually impaired people. *Applied Sciences* 9, 16 (2019), 3350.

[6] COCO Consortium. 2022. COCO - Common Objects in Context. https://cocodataset.org/#home. (2022).

[7] NVIDIA Corporation. 2024. Jetson - Embedded AI Computing Platform. https://developer.nvidia.com/embedded-computing. (2024).

[8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems* 36 (2024).

[9] eSight Corporation. 2024. eSight Eyewear. https://www.esighteyewear.com/. (2024).

[10] Be My Eyes. 2023. Be My Eyes - Helping the blind. https://www.bemyeyes.com/. (2023).

[11] Microsoft Garage. 2023. Seeing AI. https://www.microsoft.com/en-us/garage/wall-of-fame/seeing-ai/. (2023).

[12] Google. 2023a. Google Keep. https://keep.google.com/. (2023).

[13] Google. 2023b. Google Lens. https://lens.google.com/. (2023).

[14] Google. 2024. MediaPipe Hand Landmarker. https://developers.google.com/mediapipe/solutions/vision/hand_landmarker. (2024).

[15] CamFind Inc. 2023. CamFind. https://camfindapp.com/. (2023).

[16] Luxonis. 2024. Robotic Vision Made Simple. https://www.luxonis.com/. (2024).

[17] Meta. 2023. New Ray-Ban Meta Smart Glasses. https://about.fb.com/news/2023/09/new-ray-ban-meta-smart-glasses/. (2023).

[18] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–22.

[19] Halley Profita, Reem Albaghli, Leah Findlater, Paul Jaeger, and Shaun K Kane. 2016. The AT effect: how disability affects the perceived social acceptability of head-mounted display use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4884–4895.

[20] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.

[21] Jarek Reynolds, Chandra Kanth Nagesh, and Danna Gurari. 2024. Salient object detection for images taken by people with vision impairments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 8522–8531.

[22] Ray Smith. 2023. Tesseract OCR. `https://github.com/tesseract-ocr/tesseract`. (2023).

[23] Anna Syberfeldt, Oscar Danielsson, and Patrik Gustavsson. 2017. Augmented reality smart glasses in the smart factory: Product evaluation guidelines and review of available products. *IEEE Access* 5 (2017), 9118–9130.

[24] OrCam Technologies. 2024. OrCam MyEye 2. `https://www.orcam.com/en/myeye2/`. (2024).

[25] Ultralytics. 2021. YOLOv5: Real-Time Object Detection System. `https://github.com/ultralytics/yolov5`. (2021).

[26] Jacob O Wobbrock, Shaun K Kane, Krzysztof Z Gajos, Susumu Harada, and Jon Froehlich. 2011. Ability-based design: Concept, principles and examples. *ACM Transactions on Accessible Computing (TACCESS)* 3, 3 (2011), 1–27.

[27] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).